Zero-shot classification of ECG signals using a CLIP-based model

Navmeet S. Jassal, Konstantin Egorov\*, Semen BUDENNYY,2

1Laboratory X, Institute X, Department X, Organization X, City X, State XX (only USA, Canada and Australia), Country

2Laboratory X, Institute X, Department X, Organization X, City X, State XX (only USA, Canada and Australia), Country

**\* Correspondence:**Corresponding Author  
ndzhassal2@edu.hse.ru

Keywords: keyword1, keyword2, keyword3, keyword4, keyword5. (Min.5-Max. 8)

Abstract

For full guidelines please refer to [Author Guidelines](https://www.frontiersin.org/guidelines/author-guidelines)

As a primary goal, the abstract should render the general significance and conceptual advance of the work clearly accessible to a broad readership. References should not be cited in the abstract. Leave the Abstract empty if your article does not require one – please see the “Article types” on every Frontiers journal page for full details.

# Introduction

According to the World Heart Federation (WHF), a leader in global cardiovascular health, among other reports from leading clinicians, researchers and institutions, cardiovascular diseases (CVDs) are a leading cause of mortality worldwide [17; 22; 25; 36]. In the United States, the Center for Disease Control (CDC) estimates that the annual total cost associated with CVD-related treatment and mortality is approximately $219 billion dollars (USD) per year [18; 21]. Given the financial and socioeconomic burdens of CVDs, it is essential that individuals receive timely and accurate medical care for the diagnosis, treatment, and ongoing care of CVDs.

An electrocardiogram (ECG) remains the medical standard and benchmark for the identification and diagnosis of CVDs with more than 300 million ECGs being performed globally [7]. In general, ECGs measure the changes in electrical membrane potential of the heart across different directions of the body and are often referred to as leads with a 12-lead ECG report being the most common type of report [14; 29]. Once an ECG report is available, it is ready to be analyzed by a licensed medical professional such as a cardiologist or an auxiliary medical professional.

However, it is important to note that even today, interpreting and analyzing ECG reports remains a highly complex and time-consuming process for medical professionals [11]. Not only are ECG reports manually annotated but the individual analyzing the ECG report must possess a high degree of technical skills and a vast understanding of topics such as cardiac anatomy, electrophysiology, pattern recognition, coronary distribution, and pathophysiology to perform correctly and accurately [11]. In one systematic literature review by Cook et al., it was found that medical professionals, such as cardiologists, across levels of education experienced challenges in providing a correct diagnosis for an ECG report with correct interpretation accuracy ranging from 49% to 92% and with a median interpretation accuracy of 57% [11]. In the same systematic literature review, it was also found that medical professionals who received continuing education and training related to ECG interpretation and analysis, did improve their ability to correctly provide a diagnosis for an ECG report with a median accuracy of 67%, suggesting that ongoing training for medical professions can play an important role in providing correct diagnoses [11].

Seeing the challenge that medical professionals faced with ECG interpretation, many computer scientists and physicians aimed to address this problem. Namely, with the numerous advancements observed in machine learning (ML), deep learning (DL), and artificial intelligence (AI) vision models over recent years, it was postulated and believed that using ML and DL vision models may be able to correctly classify pathologies related to ECG reports [citation needed].

In this paper, our goal is to build upon and contribute to the body of work of DL literature as it relates ECG classification. Particularly, in this study, we demonstrate that using a multimodal Contrastive language image pre-training (CLIP) framework is effective in the diagnoses of any number and combination of ECG diagnostic classes. After conducting a thorough literature search and review, it is to the best of our knowledge that using a multimodal CLIP framework within the field of ECG classification has not yet been explored extensively with a only handful of studies related to this topic being published within the last two years. Overall, we show that…. [To be continued upon results]

In the next section, Related Works, an overview of CLIP and the current state of ML & DL models for ECG classification will be provided. In addition, the topic of Zero-shot classification as it relates to our study will also be discussed.

# Related Works

## Contrastive language image pre-training (CLIP)

In 2021, OpenAI researchers published their paper, “Learning Transferable Visual Models From Natural Language Supervision” which introduced a novel approach to multimodal learning, namely, Contrastive language image pre-training or CLIP for short [10; 20]. CLIP is a multimodal framework that comprises of two components, an image encoder (vision model) and a text encoder (text model) [10; 20]. When using a CLIP framework, the image encoder is trained alongside the accompanying text encoder [10]. This allows the CLIP-based model to develop a comprehensive or "deep" understanding of its image-text pair inputs [20]. To facilitate the learning process, the CLIP framework utilizes contrastive loss (also known as symmetric loss) [10; 20]. Contrastive loss calculates the distance between the embeddings of the image-text pair inputs and aims to minimize the overall distance between image-text pairs when one sample is like another (positive instance) and maximize the distance between image-text pairs inputs when samples are not similar to other samples (negative instance) [10; 20]. In doing this, the two text and image encoders are incentivized to learn better representations of their inputs to capture the semantic relationship between embeddings of image-text pairs and are penalized when they do not learn or capture reliable representations of the image-text pairs [20].

Unlike traditional ML or DL approaches for classification tasks, which are generally jointly trained with some image feature extractor and a classifier to predict the corresponding image class or label(s), the notion of using a CLIP-based model presents a novel approach for classification tasks [20]. With a CLIP-based model, the model can be utilized for zero-shot classification across a variety of classification tasks and for many types of datasets. In our study, we demonstrate that a CLIP-based model can be used within zero-shot learning (pending).

Task 1: Put Image of Contrastive Learning Pseudocode or type in LaTex and insert equations

## Related Works – Deep Learning and ECG Classification

In the fields of AI, ML, and DL, the primary goal has been to create models capable of correctly classifying CVDs from ECG reports [2]. To achieve models capable of this, a patient’s medical history including their ECG report(s) are aggregated, preprocessed, and then analyzed by the model which in the end, provides a recommendation or classification based on its input.

In the past decade, several notable ECG classification models have been developed that have demonstrated to be as effective in correctly classifying CVDs from ECGs as experienced cardiologists using ECG signal inputs [4; 23]. In one systematic literature review about state-of-the-art (SOTA) DL methods for ECG classification by Petmezas and colleagues, it was found that the most popular DL models for ECG classification tasks included variants of convolutional neural networks (CNNs), recurrent neural networks (RNNs), ResNet models, and LSTM models [33]. From the studies reviewed in this systematic review, 66.1% favored using CNNs for specific, downstream ECG classification tasks [33]. In Lu et al.’s study, the researchers achieved an accuracy of 99.31% for arrhythmia classification using a 1D-CNN [19; 33]. In a different study, authors Yu et al. achieved an accuracy of 99.70 % for premature ventricular contraction also utilizing a 1D-CNN demonstrating the robustness of CNNs for different types of, singular-class of ECG classification tasks [6; 33]. Other studies have also demonstrated that training and using CNNs is effective for CVD classification tasks for classes such as myocardial infarction, coronary artery disease, and congestive heart failure to name a few.

## Related Works – CLIP-based models in medical imaging

Using a CLIP-based model has proven to be effective across a wide range of datasets and classes such as s Food-101, CIFAR-100, ImageNet, and MNIST to name a few [20]. However, the utilization of CLIP-based models in medical imaging for image classification or diagnostic tasks remains an area of research that has not been analyzed or explored extensively [9]. In their systematic literature review on the topic of CLIP-based models for medical imaging, Zhao et al. found that since 2021, there have been approximately 38 studies that have utilized the CLIP-based model framework for medical imaging tasks [9]. From these 38 studies, some notable examples include seven studies using the CLIP-based framework for the classification of chest X-rays, one study using the CLIP-based framework for the stomach histology, one study for CT scans of the lungs, one study for the classification of skin/dermatology conditions, one study for MRIs of brains, and one study used the CLIP-based framework for eye classification tasks [9].

In Zhao et al.’s systematic review, there was only one study that used the CLIP-based framework for ECG classification tasks which used ECG signals as their primary input [9; 15]. In their paper, "ETP: LEARNING TRANSFERABLE ECG REPRESENTATIONS VIA ECG-TEXT PRE-TRAINING", authors Liu et al. describe training a CLIP-based model from scratch using the PTB-XL and CPSC2018 datasets [3; 15; 30]. In this study, 12-lead ECG reports and their corresponding ECG signals were used to train a CLIP-based model where a 1D-ResNet18 model served as the image encoder and BioClinicalBERT was used as the text encoder [15; 31]. During the training of the image encoder, the 1D-ResNet18 model had its weights updated during training whereas the text encoder, BioClinicalBERT, had its weights remain constant or frozen [15]. The aim of Liu et al.’s study was to demonstrate the viability of zero-shot ECG classification tasks [15]. In their experiments, the joint multimodal training of image and text encoders simultaneously achieved SOTA results for ECG classification on the PTB-XL and CPSC2018 datasets, respectively, during both supervised learning as well as during zero-shot classification [15]. In the PTB-XL zero-shot classification experiment, their ETP model achieved higher AUC, ACC, and F1 scores across four distinct classes of ECGs [15]. Similarly, their ETP approach also achieved significant results on the CPSC2018 dataset achieving higher AUC, ACC, and F1 scores across nine distinct classes of ECGs [15]. Both experiments demonstrated the robustness of using the CLIP-based framework for ECG classification in both supervised learning and zero-shot learning/classification [15]. Liu et al.’s study was important because it was the first study that demonstrated that the CLIP framework could be used for zero-shot classification of ECGs and in this study, we aim to build on their work across a greater number of ECG diagnostic classes and larger datasets.

Another study that demonstrates the efficacy and value of multimodal or joint training of image and text encoders was seen in Li et al.’s, "Frozen Language Model Helps ECG Zero-Shot Learning" [16]. In this study, the authors introduced a new technique called Multimodal ECG-text SSL or METS for short [16]. Like Liu et al.’s study, Li et al. in their study also use a text encoder with frozen weights where the goal of METS is to use automatically generated ECG clinical reports to guide the training of the ECG image encoder [15; 16]. The METS approach achieved an improvement of 10% in performance without the use of annotated datasets relative to comparable studies in this field [16].

In summary, the METS model takes an ECG signal and its corresponding ECG report/diagnosis text as inputs and feeds them into a multimodal learning framework, like CLIP [16; 20]. Then, using contrastive loss, they attempted to measure the distance between image-text pairs so that their ECG image encoder can learn "deep" representations of its 9 inputs [16]. When it came to evaluation, the METS approach achieved higher accuracy, precision, recall, and F1 scores in the PTB-XL compared to other leading methods [16].

## Related Works – Zero-shot Learning

* Task: Need specific articles that describe Zero-Shot Learning

According to Hugging Face, zero-shot classification is a natural language processing (NLP) task where a model is trained on a specific set of labeled image-text pair data but can generalize and classify new examples of classes that were not in the labeled image-text pair dataset during the model’s training [38]. In this study, a simple example of zero-shot classification would be if the image model were trained on a set of labeled image-text pairs with some classes of image-text pairs that were excluded during training (for example, ECG signals for myocardial infarction). Then, during the testing phase, the trained image model would be able generalize to new data and in turn be able to classify myocardial infarction.

# Materials and Methodology

## Datasets

Two experiments were performed in this study, Experiments A and B, respectively.

### Datasets – Experiment A

In Experiment A, a CLIP-based model consisting of a text encoder and image encoder, was trained on the Physikalisch-Technische Bundesanstalt (PTB-XL) dataset. The PTB-XL dataset contains ECG image-text pairs where each image is represented by a raw 10 second 12-lead ECG signal and where the ECG clinical report represents each text input. In Experiment A, the CLIP-based model was trained on most of all classes in the PTB-XL dataset with some exceptions. Those classes which were excluded during the training of the CLIP-based model were utilized for the testing and evaluation of the CLIP-based model where the CLIP-based model would encounter ECG diagnostic classes that it had not previously encountered or seen during its training. In doing this, the CLIP-based model could be evaluated in a zero-shot classification setting to determine how well the CLIP-based model is able to generalize to unseen data.

excluded\_classes = ['left ventricular hypertrophy',

'st depression',

'low qrs voltages',

's t changes',

'sinus tachycardia',

't wave abnormal',

'right axis deviation']

**Figure 1**: Distribution of Classes (labels) in the PTB-XL dataset

|  |  |
| --- | --- |
| **Diagnostic Class** | **Total Samples** |
| sinus rhythm | 18058 |
| myocardial infarction | 5248 |
| left axis deviation | 5146 |
| abnormal QRS | 3389 |
| left ventricular hypertrophy | 2354 |
| t wave abnormal | 2341 |
| myocardial ischemia | 2171 |
| left anterior fascicular block | 1624 |
| atrial fibrillation | 1514 |
| ventricular ectopics | 1153 |
| incomplete right bundle branch block | 1118 |
| st depression | 1009 |
| sinus tachycardia | 826 |
| 1st degree av block | 795 |
| nonspecific intraventricular conduction disorder | 787 |
| sinus arrhythmia | 772 |
| s t changes | 768 |
| sinus bradycardia | 637 |
| qwave abnormal | 548 |
| complete right bundle branch block | 542 |
| left bundle branch block | 536 |
| left atrial enlargement | 426 |
| premature atrial contraction | 398 |
| nonspecific st t abnormality | 381 |
| anterior myocardial infarction | 353 |
| right axis deviation | 343 |
| prolonged pr interval | 340 |
| t wave inversion | 294 |
| pacing rhythm | 294 |
| inferior ischaemia | 218 |
| low qrs voltages | 182 |
| left posterior fascicular block | 177 |
| supraventricular premature beats | 157 |
| indeterminate cardiac axis | 155 |
| lateral ischaemia | 140 |
| right ventricular hypertrophy | 126 |
| prolonged qt interval | 118 |
| right atrial hypertrophy | 99 |
| ventricular bigeminy | 82 |
| wolff parkinson white pattern | 79 |
| incomplete left bundle branch block | 77 |
| atrial flutter | 73 |
| anterior ischemia | 44 |
| ventricular hypertrophy | 29 |
| st elevation | 28 |
| supraventricular tachycardia | 27 |
| paroxysmal supraventricular tachycardia | 24 |
| ventricular trigeminy | 20 |
| complete heart block | 16 |
| 2nd degree av block | 14 |

### Datasets – Experiment B

In Experiment B, the primary dataset used was the [insert dataset here, subject to change].

## Data pre-processing

In Experiments A and B, the raw ECG signal for each ECG image-text pair was down sampled from its original frequency to 128 Hz. After down sampling the ECG signals, the ECG signals were then normalized by calculating the average mean and average standard deviation for all samples. The down sampled dataset is then filtered to exclude any diagnostic classes with fewer than 100 samples to ensure that sufficient data was available to train the models in both experiments.

## Model architecture

### Text Encoder – BioClinicalBERT

In Experiment A, the text encoder used as part of the CLIP-based model was BioClinicalBERT. BioClinicalBERT is a variant of the BioBERT base model which is additionally trained on all notes from MIMIC III, a database of electronic health records from ICU patients at Beth Israel Hospital in Boston, MA and which contains over 880 million words [31]. The pretrained BioClinicalBERT is publicly available through HuggingFace.

The implementation of the BioClinicalBERT text encoder in Python3 can be seen in the Figure below.

* Task: For LaTex, insert code correctly within a code block.

### Image Encoder – 1D-CNN or [SOME OTHER MODEL, TBD]

## Model training

In Experiment A, the CLIP-based model consisting of a 1D-CNN image encoder and the BioClinicalBERT text encoder was trained and validated on the PTB-XL dataset. The CLIP-based model was trained over 20 epochs with a learning rate of 0.001, Adam optimizer, a StepLR scheduler,

|  |  |
| --- | --- |
| **Hyperparameter Type** | **Value** |
| Epochs | 20 |
| Optimizer | torch.optim.Adam |
| Projection Dimension | 64/128/256/512/768/1536 |
| Projection Head Learning Rate | 0.001 |
| Image Encoder Learning Rate | 0.001 |
| Image Encoder Type | 1D-CNN |
| Text Encoder Type | BioClinicalBERT |
| Temperature | 1.0 |
| ECG Signal Sample Rate (Hz) | 128 |
| Dropout | 0.0 |
| Criterion (Loss Function) | Contrastive Loss |

* Task: Create LaTex table for hyperparameters used

# Results

### 4.1 Results – Experiment A

### 4.2 Results – Experiment B

# Discussion

# References

1. A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. / H. Liu [et al.] // Sci Data. — 2022. — Vol. 9. — DOI: https://doi.org/10.1038/s41597-022-01403-5.

2. Ahsan M.M. Luna A.S. S. Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review // Healthcare (Basel, Switzerland). — 2022. — Vol. 10, no. 3. — P. 1461–1471. — PMID: 35327018.

3. An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection / F. Liu [et al.]. — 2018. — Journal of Medical Imaging and Health Informatics.

4. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management / C. K. Siontis [et al.] // Nature Reviews Cardiology. — 2021. — Vol. 18. — P. 465–478. — DOI: 10.1038/s41569-020- 00503-2.

5. Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study / H. Zhu [et al.] // The Lancet. Digital health. — 2020. — Vol. 2, no. 7. — P. 348–357. — DOI: 10.1016/S2589-7500(20)30107-2.

6. Automatic Premature Ventricular Contraction Detection Using Deep Metric Learning and KNN / J. Yu [et al.] // Biosensors. — 2021. — Vol. 11, no. 3. — DOI: ttps://doi.org/10.3390/bios11030069.

7. Chhabra L. S. Y. Electrocardiogram // StatPearls [Internet]. — 2023.

8. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020 / P. Alday [et al.] // Institute of Physics and Engineering in Medicine. — 2020. — Vol. 41, no. 12. — DOI: <https://doi.org/10.1088/1361-6579/abc960>.

9. CLIP in Medical Imaging: A Comprehensive Survey / Z. Zhao [et al.] // arXiv. — 2023. — DOI: arXiv:2312.07353.

10. CLIP: Connecting text and images / A. Radford [et al.]. — 2021.

11. Cook A.D. Oh S.Y. P. V. Accuracy of Physicians’ Electrocardiogram Interpretations: A Systematic Review and Meta-analysis // JAMA internal medicine. — 2020. — Vol. 180, no. 11. — P. 1461–1471. — PMID: 32986084.

12. Diagnostic statement dictionary / H. Liu [et al.]. — 05/2022. — DOI: 10.6084/m9.figshare.17912507. v1. — URL: https://springernature.figshare.com/articles/dataset/diagnostic\_statement\_ dictionary/17912507 ; 12-lead electrocardiogram; ECG; Multi-label; ECG abnormalities. Figshare.

13. ECG records / H. Liu [et al.]. — 2022. — DOI: 10 . 6084 / m9 . figshare . 17912444 . v1. — URL: https://doi.org/10.6084/m9.figshare.17912444.v1. Figshare.

14. Electrocardiography: A Technologist’s Guide to Interpretation / C. Tso [et al.] // Journal of Nuclear Medicine Technology. — 2015. — Vol. 43, no. 4. — P. 247–252. — DOI: 10.2967/jnmt.115.163501.

15. ETP: LEARNING TRANSFERABLE ECG REPRESENTATIONS VIA ECG-TEXT PRE-TRAINING / C. Liu [et al.] // arXiv. — 2023. — DOI: arXiv:2309.07145v1.

16. Frozen Language Model Helps ECG Zero-Shot Learning / J. Li [et al.] // arXiv. — 2023. — DOI: arXiv:2303.12311.

17. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the Global Burden of Disease Study 2016 // THE LANCET. — 2017. — Vol. 390, 10100. — P. 1151–1210. — PMID: 28919116.

18. Heart Disease and Heart Attack. — Centers for Disease Control, Prevention, 2021. — URL: https: //www.cdc.gov/policy/polaris/healthtopics/heartdisease/index.html ; Accessed: 2024-01-15. [Electronic resource].

19. KecNet: A Light Neural Network for Arrhythmia Classification Based on Knowledge Reinforcement / P. Lu [et al.] // Journal of Healthcare Engineering. — 2021. — DOI: https://doi.org/10.1155/2021/ 6684954. — eprint: 6684954.

20. Learning Transferable Visual Models From Natural Language Supervision / A. Radford [et al.] // Proceedings of Machine Learning Research. — 2021. — Vol. 139. — P. 8748–8763.

21. Lui J.N. et al. Impact of New Cardiovascular Events on Quality of Life and Hospital Costs in People With Cardiovascular Disease in the United Kingdom and United States // Journal of the American Heart Association. — 2023. — Vol. 12, no. 19. — ISSN e030766. — DOI: 10.1161/JAHA.123.030766. — URL: <https://www.ahajournals.org/doi/10.1161/JAHA.123.030766>.

22. Machine learning-based detection of cardiovascular disease using ECG signals: performance vs. complexity / H. Pham [et al.] // Front. Cardiovasc. Med. — 2023. — Vol. 10. — DOI: 10.3389/fcvm.2023. 1229743. 20

23. Mamun K.R.M.M. E. T. AI-Enabled Electrocardiogram Analysis for Disease Diagnosis // Applied System Innovation. — 2023. — Vol. 6, no. 5. — DOI: <https://doi.org/10.3390/asi6050095>.

24. Mapping from Chinese ECG statements to AHA codes / H. Liu [et al.]. — 05/2022. — DOI: 10.6084/ m9.figshare.19738468.v1. — URL: https://springernature.figshare.com/articles/dataset/ Mapping\_from\_Chinese\_ECG\_statements\_to\_AHA\_codes/19738468 ; 12-lead electrocardiogram; ECG; Multi-label; ECG abnormalities. Figshare.

25. Mensah A.G. Roth A.G. F. V. The Global Burden of Cardiovascular Diseases and Risk Factors: 2020 and Beyond // Journal of the American College of Cardiology. — 2019. — Vol. 74, no. 20. — P. 2529–2532. — ISSN 0735-1097. — DOI: 10.1016/j.jacc.2019.10.009. — URL: https://www.sciencedirect.com/ science/article/pii/S0735109719379288.

26. OpenStax College Anatomy & Physiology //. — OpenStax College, 2013. — URL: https://commons. wikimedia.org/w/index.php?curid=30148227.

27. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals / A. Goldberger [et al.] // Circulation [Online]. — 2000. — Vol. 101, no. 23. — P. 215–220.

28. Pollock D., Makaryus N. Physiology, Cardiac Cycle // StatPearls. — 2022.

29. Prutkin M. ECG tutorial: Basic principles of ECG analysis // UpToDate. — 2023.

30. PTB-XL, a large publicly available electrocardiography dataset / P. Wagner [et al.]. — 2020. — URL: https://www.nature.com/articles/s41597-020-0495-6. Scientific Data.

31. Publicly Available Clinical BERT Embeddings / E. Alsentzer [et al.] // Proceedings of the 2nd Clinical Natural Language Processing Workshop. — Association for Computational Linguistics, 2019. — P. 72– 78. — DOI: 10.18653/v1/W19-1909. — URL: <https://www.aclweb.org/anthology/W19-1909>.

32. Scikit-learn: Machine Learning in Python / F. Pedregosa [et al.] // Journal of Machine Learning Research. — 2011. — Vol. 12. — P. 2825–2830.

33. State-of-the-Art Deep Learning Methods on Electrocardiogram Data: Systematic Review / G. Petmezas [et al.] // JMIR Medical Informatics. — 2022. — Vol. 10, no. 8. — PMID: 35327018.

34. The attributes of ECG records / H. Liu [et al.]. — 2022. — DOI: 10.6084/m9.figshare.17912441.v1. — URL: https://doi.org/10.6084/m9.figshare.17912441.v1. Figshare.

35. Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021 / M. Reyna [et al.] // PhysioNet. — 2022. — DOI: <https://doi.org/10.13026/34va7q14>.

36. World Heart Report 2023. — World Heart Federation, 2023. — URL: https : / / world - heart - federation.org/resource/world-heart-report-2023/ ; Accessed: 2023-10-15. [Electronic resource].

37. Yanowitz G. 1. The Standard 12 Lead ECG // ECG LEARNING CENTER. —.

38. Zero-Shot Classification. — 2023. — URL: https : / / huggingface . co / tasks / zero - shot - classification